ShortGPT: Layers in Large Language Models are More Redundant Than You Expect

Year: 2024 | *Citations: 230* | *Authors: Xin Men, Mingyu Xu, Qingyu Zhang*

Abstract

As Large Language Models (LLMs) continue to advance in performance, their size has escalated significantly, with current LLMs containing billions or even trillions of parameters. However, in this study, we discovered that many layers of LLMs exhibit high similarity, and some layers play a negligible role in network functionality. Based on this observation, we define a metric called Block Influence (BI) to gauge the significance of each layer in LLMs. We then propose a straightforward pruning approach: layer removal, in which we directly delete the redundant layers in LLMs based on their BI scores. Experiments demonstrate that our method, which we call ShortGPT, significantly outperforms previous state-of-the-art (SOTA) methods in model pruning. Moreover, ShortGPT is orthogonal to quantization-like methods, enabling further reduction in parameters and computation. The ability to achieve better results through simple layer removal, as opposed to more complex pruning techniques, suggests a high degree of redundancy in the model architecture.